

Designing A Composite Dictionary Adaptively From Joint Examples

Zhangyang Wang[†], Yingzhen Yang[†], Jianchao Yang[‡], Thomas Huang[†]

[†]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[‡]Snapchat Inc, Venice, CA 90291, USA

{zwang119, yyang58}@illinois.edu, jianchao.yang@snapchat.com, t-huang1@illinois.edu

Abstract—We study the complementary behaviors of external and internal examples in image restoration, and are motivated to formulate a composite dictionary design framework. The composite dictionary consists of the global part learned from external examples, and the sample-specific part learned from internal examples. The dictionary atoms in both parts are further adaptively weighted to emphasize their model statistics. Experiments demonstrate that the joint utilization of external and internal examples leads to substantial improvements, with successful applications in image denoising and super resolution.

I. INTRODUCTION

Sparse coding (SC) by representing a signal as a sparse linear combination of representation bases (dictionary atoms) has been widely applied [1]. The dictionary, which should both faithfully represent the signal and effectively extract task-specific features, plays an important role [2]. For image restoration, classical methods either rely on a large external set of image examples [3], or find self-similar examples from the input [4]. With much progress being made, it is recently recognized that external and internal examples each suffer from certain drawbacks, but their properties may be complementary [5], [6].

We believe the joint utilization of external and internal examples in dictionary design is crucial for further improving image restoration. We thus formulate a new composite dictionary design framework for image restoration tasks. Successful applications in image denoising and super resolution (SR) demonstrate its effectiveness.

II. RELATED WORK

The problem to be investigated resembles to a general problem in SC-based classification: how to adaptively build the relationship between dictionary atoms and class labels? Based on predefined relationships, current supervised dictionary learning (DL) methods are categorized into either learning a shared dictionary by all classes, which may be compact but not sufficiently discriminative [7]; or a class-specific dictionary

with the opposite properties [8]. In [9], the authors jointly learned a composite dictionary combining class-specific and shared dictionary atoms, with a latent matrix indicating the relationship between dictionary atoms and labels.

In analogy to the classification case, reconstruction-purpose dictionaries have been built from either external or internal examples. External example-based methods are known for their capabilities to produce plausible image appearances. However, there is no guarantee that an arbitrary input patch can be well matched or represented by a pre-fixed external set. When there is rarely any match for the input, external examples are prone to introduce either noise or oversmoothness [10]. Meanwhile, the self similarity property supplies internal examples that are highly relevant to the input, but only of a limited number. Due to the insufficiency of internal examples, their mismatches often result in severe visual artifacts [11].

The joint utilization of both external and self examples has been first studied for image denoising [12]. Mosseri et. al. [13] proposed that image patches had different preferences towards either external or self examples for denoising. Such a preference is in essence the tradeoff between noise-fitting versus signal-fitting. In [5], [6], a joint super-resolution (SR) models was proposed to adaptively combine the advantages of both external and self example-based loss functions. [14] further investigated the utilization of self-similarity into deep learning-based SR. However, none of the prior work makes much progress towards a unified dictionary design framework.

III. TECHNICAL APPROACH

A. Overview

The *composite dictionary* consists of the *global dictionary* part learned from external examples, and the *sample-specific dictionary* part learned from internal examples. The atoms in both parts are further weighted to exploit the different model statistics. Given input signal $\mathbf{x} \in R^{p \times 1}$, the formulation can be mathematically presented as:

$$\min_{\mathbf{a}, \omega_G, \omega_S} \lambda_E \sum_i \|\mathbf{a}_i^G\|_1 + \lambda_I \sum_j \|\mathbf{a}_j^S\|_1 + \|\mathbf{x} - \sum_i \mathbf{d}_i^G M_G(\mathbf{d}_i^G, \mathbf{x}, \omega_G) \mathbf{a}_i^G - \sum_j \mathbf{d}_j^S M_I(\mathbf{d}_j^S, \mathbf{x}, \omega_S) \mathbf{a}_j^S\|_F^2 \quad (1)$$

Here $\mathbf{d}_i^G \in R^{p \times 1}, i = 1, 2, \dots, M$ denotes the dictionary atoms pre-learned from external examples, and $\mathbf{d}_j^S \in$

$R^{p \times 1}, j = 1, 2, \dots, N$ the atoms pre-learned from internal examples. We define $\mathbf{D}^G = \{\mathbf{d}_1^G, \dots, \mathbf{d}_M^G\}$ as the *global base dictionary* and $\mathbf{D}^S = \{\mathbf{d}_1^S, \dots, \mathbf{d}_N^S\}$ as the *sample-specific base dictionary*. $\mathbf{a} \in R^{(M+N) \times 1}$ denotes sparse codes of \mathbf{x} , consisting of $\mathbf{a}_i^G, i = 1, 2, \dots, M$ and $\mathbf{a}_j^S, j = 1, 2, \dots, N$, corresponding to \mathbf{D}^G and \mathbf{D}^S , respectively. λ_E and λ_I are constants. Note that both the first two terms, and the last two summations in the third term of (1), can be each combined together just like in conventional SC. We prefer writing them separately in a purpose to highlight two different dictionary parts. M_G and M_I denote some similarity-based weights between the dictionary atoms and the input, parametrized by ω_G and ω_S , respectively.

Our solution to (1) takes three steps:

- Obtain \mathbf{D}^G and \mathbf{D}^S prior to solving (1).
- Choose the desired forms of M_G and M_I .
- Solve (1) using a coordinate-descent algorithm.

Note the strategy is to first fix base dictionaries, then adapting them to the input by learning weights, which is close to [15]. Both lead to efficient and flexible dictionary representations. Yet rather than simply enforcing sparsity constraints on the weights [15], we aim to build a more adaptive relationship between the input and the atoms, based on the complementary example statistics. Additionally, while the base in [15] is simply a DCT dictionary, our \mathbf{D}^G and \mathbf{D}^S are specifically crafted from external and internal examples separately.

Algorithm 1 Coordinate descent algorithm for solving (1)

Require: \mathbf{D}^G and \mathbf{D}^S ; λ_E and λ_I ; $ITER$; β .

- 1: FOR $t=1$ to $ITER$ DO
- 2: Fix Ω_G and Ω_S , solve (1) over \mathbf{A} using the feature-sign algorithm [16].
- 3: Fix Ω_G and \mathbf{A} , solve Ω_S by taking gradient descent over \mathbf{F}_S , with step size β .
- 4: Fix Ω_S and \mathbf{A} , solve Ω_G by taking gradient descent over \mathbf{F}_G , with step size β .
- 5: END FOR

Ensure: \mathbf{A} , Ω_G and Ω_S

B. Algorithm

First of all, we learn \mathbf{D}^G and \mathbf{D}^S from the sets of external and internal examples respectively (e.g., by K-SVD [1]). Before moving on to learn functional forms of M_G and M_I , it is interesting to examine whether a static, but well-defined weight could help. Without loss of generality, we assume both M_G and M_I have a value range between $[0,1]$. It is obvious that, when M_G (or M_I) becomes larger, i.e., the current atom is highly correlated to the input, the atom will be more favored by the ℓ_1 penalty. We first define M_G and M_I both in the form of radial basis function (RBF) kernels:

$$\begin{aligned} M_G(\mathbf{d}_i^G, \mathbf{x}, \omega_G) &= \exp(-\omega_G \|\mathbf{d}_i^G - \mathbf{x}\|^2) \\ M_I(\mathbf{d}_j^S, \mathbf{x}, \omega_S) &= \exp(-\omega_S \|\mathbf{d}_j^S - \mathbf{x}\|^2) \end{aligned} \quad (2)$$

where ω_G and ω_S are both fixed constants, but of different values. As discussed above, it is often more likely to find

“highly similar” examples internally than from external examples. On the other hand, the external set can usually provide more abundant “reasonably similar” (not necessarily highly though) examples. We thus desire the value of M_I to decrease more quickly than M_G , which means ω_S is supposed to be chosen larger than ω_G . Note that when ω_G and ω_S become fixed, (1) becomes a plain sparse decomposition problem, that can be solved efficiently [16].

Inspired by (2), it is straightforward to construct parameterized M_G and M_I in the form of Mahalanobis kernel [17]:

$$\begin{aligned} M_G(\mathbf{d}_i^G, \mathbf{x}, \Omega_G) &= \exp(-(\mathbf{d}_i^G - \mathbf{x})^T \Omega_G (\mathbf{d}_i^G - \mathbf{x})) \\ M_I(\mathbf{d}_j^S, \mathbf{x}, \Omega_S) &= \exp(-(\mathbf{d}_j^S - \mathbf{x})^T \Omega_S (\mathbf{d}_j^S - \mathbf{x})) \end{aligned} \quad (3)$$

Note Ω_G and Ω_S are both semi-definite real matrices. However, learning Ω_G or Ω_S directly requires enforcing a positive semi-definite constraint during optimization, which is expensive. A cheaper and well-known solution is to decompose: $\Omega_G = \mathbf{F}_G^T \mathbf{F}_G$, $\Omega_S = \mathbf{F}_S^T \mathbf{F}_S$. Since \mathbf{F}_G is an unconstrained real matrix, we can now cast the metric learning as an unconstrained matrix optimization problem.

Concluding all above, we solve (1) by a coordinate descent algorithm, as detailed in Algorithm 1.

IV. EXPERIMENTS

In our experiments, \mathbf{x} is by default a 5×5 image patch, columnized to be a 25×1 vector; the external/internal examples and resulting dictionary atoms share the same dimension. We use the natural patches cropped from the Berkley Segmentation Dataset (BSD) as the collection of external examples. The internal example candidates are cropped from the input image with a spatial overlap of 1, to ensure a reasonably sufficient amount. For image SR, the methodology is similar but works on example pairs. We set $M = 128$, and $N = 32$ as the default dictionary sizes of \mathbf{D}^G and \mathbf{D}^S .

In Algorithm 1, \mathbf{F}_G and \mathbf{F}_S are both initialized to be identity matrices with some random disturbance on the diagonal elements. When handling (2), we fine-tune ω_G and ω_S by cross-validation. The solved \mathbf{A} from (2) serves as a proper initialization in (3). λ_E and λ_I vary by applications and will be tuned, but we find it universally applicable to set λ_I around 10 times of λ_E . β is fixed at 0.9, and $ITER$ is 5 for all. When well initialized, the current MATLAB implementation takes no more than 5 iterations to converge, and each iteration consumes 10-15 minutes for a 256×256 image.

A. KNN versus KSVD: The Power of Weights

As the choice of base dictionaries can be quite flexible per application requirements, it is interesting to evaluate if we could rely on simple, computationally cheap bases for a comparable performance to the sophisticated ones.

We construct the “KNN base dictionaries”: for \mathbf{D}^G , we use a K-NN clustering over the external examples and M cluster centroids are obtained. For \mathbf{D}^S , we simply obtain N closest internal examples from the input image. We then compare the following methods in a typical image denoising setting:

- **Method I.** Solving (1) using \mathbf{D}^G and \mathbf{D}^S

- **Method II.** Performing conventional SC over the composite dictionary of \mathbf{D}^G and \mathbf{D}^S (equivalent to let $M_G = M_I = \mathbf{I}$), as a benchmark.
- **Method III.** The K-SVD algorithm [1] is first applied to either external or internal examples, to obtain the global and sample-specific K-SVD dictionaries, respectively. The two K-SVD dictionaries are concatenated into a composite dictionary, over which SC is performed.

Five natural images, *Lena*, *Barbara*, *Boats*, *House* and *Peppers* are used for testing, with gaussian noise of standard deviation $\sigma = 10$. As in Table I, it is impressive to see that Method I, which relies on re-weighting the simplest KNN base dictionaries in (1), outperforms the canonical KSVD dictionaries. We also see a large average margin of 4dB of Method I over Method II in terms of PSNR, which clearly manifests the benefits of modeling and learning proper weights.

TABLE I
COMPARISON OF PNSRS (DB) USED THREE DIFFERENT METHODS.

	<i>Lena</i>	<i>Barbara</i>	<i>Boats</i>	<i>House</i>	<i>Peppers</i>
Method I	35.57	33.98	33.83	33.56	34.93
Method II	31.21	30.41	31.24	29.43	30.67
Method III	35.36	34.24	33.62	34.76	34.32

B. Application I: Image Denoising

Image denoising is a most classical application scenario for SC and DL. Each image is processed in a patch-wise manner with a spatial overlap of 1. For the best performances, we use the global and sample-specific K-SVD dictionaries, obtained in the above Method III, as our base dictionaries \mathbf{D}^G and \mathbf{D}^S . We compare the following methods on the five natural images:

- **KSVD G** denotes SC directly performed over the global K-SVD dictionary \mathbf{D}^G .
- **KSVD S** denotes SC directly performed over the sample-specific K-SVD dictionary \mathbf{D}^S .
- **KSVD C** denotes SC directly performed over the composite dictionary of \mathbf{D}^G and \mathbf{D}^S .
- **SC FW** denotes “SC with fixed weights” by solving (1) over the composite dictionary of \mathbf{D}^G and \mathbf{D}^S , with M_G and M_I defined in (2).
- **SC LW** denotes “SC with learned weights” by solving (1) over the composite dictionary of \mathbf{D}^G and \mathbf{D}^S , with M_G and M_I defined in (3).

σ varies from 10 to 50, with a stride of 10. For each method, the *average* PNSRs over all five images under various σ values are reported in Table II. The proposed SC LW outperforms all else with a large margin of around 2dB, in all cases.

The learned weights help SC LW outperform SC FW remarkably. When σ goes larger, the performances of SC FW degrade quickly. It can be interpreted that the (static) RBF weights become less reliable in describing the correlations between the noisy input patch and dictionary atoms, especially those from \mathbf{D}^G which are cropped from noise-free images. In contrast, the learned weights show better robustness. Also, it is

TABLE II
COMPARISON OF PNSRS (DB) AMONG DIFFERENT DENOISING METHODS.

	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	$\sigma = 40$	$\sigma = 50$
KSVD G	33.57	30.18	28.83	26.43	25.32
KSVD S	34.23	31.02	28.94	26.66	25.48
KSVD C	34.46	32.24	29.62	26.76	25.67
SC FW	34.83	33.45	30.28	26.27	25.32
SC LW	36.27	34.24	32.83	28.76	26.32

not a surprise to see that the utilization of joint examples leads to the consistent superiority of KSVD C over either KSVD G or KSVD S.

Interestingly, by comparing KSVD S and KSVD G, we observe that internal examples gain advantages over external ones under small σ s. For large σ s, the results of KSVD S deteriorate faster and become worse than KSVD G when $\sigma = 50$. The performance margin of KSVD C over KSVD G is also reduced when σ increases. Such observations imply that when the noise becomes heavy, internal examples are overly corrupted and cannot provide relevant references well. That coincide with the conclusion in [1], and further inspires us to investigate the ratio of \mathbf{D}^G size to \mathbf{D}^S size, denoted as r .

External versus Internal Table III is one more set of convincing results to demonstrate the complementary behaviors of joint examples. Provided the total amount of dictionary atoms is fixed at 160, Table III lists how the average PSNR of SC LW changes with r , where \mathbf{D}^G has $160r/(1+r)$ atoms and \mathbf{D}^S has $160/(1+r)$ atoms (previously $r = 4$). As shown by Table III, increasing r from 4 to 7 leads to improved PNSRs in heavy noise cases ($\sigma = 40$ and 50). However, neither an overly large nor a small r leads to any performance gain. On the one hand, the PSNR decreases rapidly with r when $r \leq 4$, under all σ s. It proves the key role of external examples in reconstructing high-quality patches under mild noise conditions. On the other hand, the PSNR also becomes poor when $r = 15$. The study of r offers another powerful support for the importance of learning composite dictionaries from joint examples.

TABLE III
COMPARISON OF PNSRS (DB) OF SC LW UNDER DIFFERENT r AND σ VALUES.

	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	$\sigma = 40$	$\sigma = 50$
$r=0$	30.47	28.48	27.67	26.20	24.06
$r=1$	32.23	30.27	29.95	26.82	25.03
$r=3$	35.46	33.27	31.80	28.97	26.21
$r=4$	36.27	34.24	32.83	28.76	26.32
$r=7$	36.18	34.05	32.58	28.94	26.37
$r=9$	36.07	33.84	31.73	28.31	26.02
$r=15$	34.57	31.28	30.80	27.67	25.83

C. Application II: Image SR

The proposed method can be applied to solving image SR problems by a variant extension. First of all, the example pools are no longer collection of image patches, but instead *example pairs* of a high-resolution (HR) patch and its low-resolution (LR) counterpart for each. Coupled dictionary learning has

been proposed in [18] to learn a *dictionary pair* from a large external set of example pairs. To be formulated mathematically, the HR and LR patch spaces $\{\mathbf{X}_{ij}\}$ and $\{\mathbf{Y}_{ij}\}$ are assumed to be tied by some mapping function. With a well-trained coupled dictionary pair $(\mathbf{D}_h, \mathbf{D}_l)$, it assumes that $(\mathbf{X}_{ij}, \mathbf{Y}_{ij})$ tends to admit a common sparse representation \mathbf{a}_{ij} . Yang et. al. [18] suggested to first infer the sparse code \mathbf{a}_{ij}^L of \mathbf{Y}_{ij} with respect to \mathbf{D}_l , and then use it as an approximation of \mathbf{a}_{ij}^H (the sparse code of \mathbf{X}_{ij} with respect to \mathbf{D}_h), in order to recover $\mathbf{X}_{ij} \approx \mathbf{D}_h \mathbf{a}_{ij}^L$.

We construct the pool of external example pairs in the same way as [18]. The pool of internal example pairs are less straightforward to construct, since the "groundtruth" HR image of the LR input is not available. To overcome the difficulty, we come up with an idea inspired by [4]. Based on the observation that singular features like edges and corners in small patches tend to repeat almost identically across different image scales, Freedman and Fattal [4] applied the "high frequency transfer" method to search the high-frequency component for a target HR patch, by NN patch matching across scales. Defining a linear interpolation operator \mathcal{U} and a downsampling operator \mathcal{D} , for the input LR image \mathbf{Y} , we first obtain its initial upsampled image $\mathbf{X}'^E = \mathcal{U}(\mathbf{Y})$, and a smoothed input image $\mathbf{Y}' = \mathcal{D}(\mathcal{U}(\mathbf{Y}))$. Given the smoothed patch $\mathbf{X}_{ij}'^E$, the missing high-frequency band of each unknown patch \mathbf{X}_{ij}^E is predicted by first solving a NN matching (4):

$$(m, n) = \arg \min_{(m,n) \in \mathcal{W}_{ij}} \|\mathbf{Y}'_{mn} - \mathbf{X}_{ij}'^E\|_F^2, \quad (4)$$

where \mathcal{W}_{ij} is defined as a small local searching window on image \mathbf{Y}' . With the co-located patch \mathbf{Y}_{mn} from \mathbf{Y} , the high-frequency band $\mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$ is pasted onto $\mathbf{X}_{ij}'^E$, i.e., $\mathbf{X}_{ij}^E = \mathbf{X}_{ij}'^E + \mathbf{Y}_{mn} - \mathbf{Y}'_{mn}$. Following this way, for the (i, j) -th patch of LR input \mathbf{Y} , we could treat \mathbf{X}_{ij}^E as its corresponding HR patch and make them an internal example pair.

TABLE IV
COMPARISON OF PNSRS (DB) USED DIFFERENT SR METHODS.

	<i>Temple</i>	<i>Train</i>	<i>Leopard</i>
Bicubic	25.29	26.14	24.14
Yang et.al.[18]	26.20	26.58	25.32
Freedman and Fattal [4]	21.17	22.54	23.04
Proposed	26.86	27.44	25.62

We then apply the coupled dictionary learning algorithm in the similar manner (with the same K , M and N) as using K-SVD above, obtaining the *global dictionary pair* $(\mathbf{D}_h^G, \mathbf{D}_l^G)$ and the *sample-specific dictionary pair* $(\mathbf{D}_h^S, \mathbf{D}_l^S)$. We solve (1) over \mathbf{D}_l^G and \mathbf{D}_l^S , after which we export the sparse codes as well as the learned weight values, to reconstruct the final HR patches by \mathbf{D}_h^G and \mathbf{D}_h^S . Comparison experiments are conducted against bicubic interpolation, Yang et.al.'s external example-based SR method [18], and Freedman et. al.'s internal example-based method [4], on three test images *Temple*, *Train* and *Leopard*, with a factor of 3. While our method is not specifically optimized for image SR, it obtains better SR results than the other two competitive methods [18], [4].

V. CONCLUSION

We propose a novel composite dictionary design framework. The composite dictionary consists of global and sample-specific parts, learned from external and internal examples, respectively. We formulate the similarity weights that adaptively correlate sparse codes with base dictionary atoms. Experiments demonstrate that the joint utilization of external and internal examples outperforms either stand-alone alternative. The applications in image denoising and SR show great potential along this research line.

REFERENCES

- [1] Michael Elad and Michal Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [2] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.
- [3] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [4] Gilad Freedman and Raanan Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 12, 2011.
- [5] Zhangyang Wang, Zhaowen Wang, Shiyu Chang, Jianchao Yang, and Thomas Huang, "A joint perspective towards image super-resolution: Unifying external-and self-examples," in *Applications of Computer Vision (WACV), Winter Conference on. IEEE*, 2014, pp. 596–603.
- [6] Zhangyang Wang, Yingzhen Yang, Zhaowen Wang, Shiyu Chang, Jianchao Yang, and T.S. Huang, "Learning super-resolution jointly from external and internal examples," *Image Processing, IEEE Transactions on*, vol. 24, no. 11, pp. 4359–4371, Nov 2015.
- [7] Julien Mairal, Francis Bach, and Jean Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, 2012.
- [8] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, "Discriminative learned dictionaries for local image analysis," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE*, 2008, pp. 1–8.
- [9] Meng Yang, Dengxin Dai, Lilin Shen, and Luc Van Gool, "Latent dictionary learning for sparse representation based classification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE*, 2014, pp. 4138–4145.
- [10] Chih-Yuan Yang, Jia-Bin Huang, and Ming-Hsuan Yang, "Exploiting self-similarities for single frame super-resolution," in *Computer Vision–ACCV 2010*, pp. 497–510. Springer, 2011.
- [11] Priyam Chatterjee and Peyman Milanfar, "Is denoising dead?," *Image Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 895–911, 2010.
- [12] Maria Zontak and Michal Irani, "Internal statistics of a single natural image," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE*, 2011, pp. 977–984.
- [13] Inbar Mosseri, Maria Zontak, and Michal Irani, "Combining the power of internal and external denoising," in *Computational Photography (ICCP), International Conference on. IEEE*, 2013, pp. 1–9.
- [14] Zhangyang Wang, Yingzhen Yang, Zhaowen Wang, Shiyu Chang, Wei Han, Jianchao Yang, and Thomas Huang, "Self-tuned deep super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–8.
- [15] Ron Rubinstein, Michael Zibulevsky, and Michael Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 1553–1564, 2010.
- [16] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng, "Efficient sparse coding algorithms," in *Advances in neural information processing systems*, 2006, pp. 801–808.
- [17] Yuya Kamada and Shigeo Abe, "Support vector regression using mahalanobis kernels," in *Artificial Neural Networks in Pattern Recognition*, pp. 144–152. Springer, 2006.
- [18] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang, "Coupled dictionary training for image super-resolution," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3467–3478, 2012.